National Statistical Service

Confidentiality Information Sheet 5

> As the world becomes more complex and computing capabilities increasingly advance, the role of microdata in statistics has become much more important. Decision makers are increasingly turning to and requesting access to microdata.
>
> Statistical Journal of the International Association for Official Statistics 26 (2009/2010) 57-63

Confidentiality

Managing the risk of disclosure in the release of microdata

Microdata are unit record data where each record represents observations for a person or an organisation. Microdata contain individual responses to questions on survey questionnaires, or administrative forms, including identifying information such as name, address, telephone number and age.

Microdata are a valuable resource for researchers and policy makers. The challenge for data custodians is striking the right balance between fulfilling obligations to protect the identity of individuals and organisations, and maximising the information available for statistical and research purposes. This requires careful weighing of the identification risks and benefits.

Types of disclosure risk in microdata

There are two key types of disclosure risk associated with microdata:

- risk that identification is made without any deliberate attempt to identify a person or organisation (spontaneous recognition); and
- risk associated with a deliberate (malicious) attempt to identify a person or organisation.

Spontaneous recognition, an identification made without any deliberate attempt, can occur if individuals with rare characteristics are present in the data. The identification risk this poses depends on how remarkable the characteristic(s) are. For example, the dataset may include people with unusual jobs (e.g. pop star or judge) or very large incomes which are highly visible in the data and could lead to their identification.

Deliberate attempts to identify a person or an organisation in a dataset may include, for example, list matching (matching unique records to external files using a combination of characteristics common to both datasets) or a 'record attack' (where a user tries to find a particular person or organisation with a set of characteristics known to the user).

Managing the risks of identification

The risks associated with providing access to microdata can be mitigated in a number of ways including:

- treating the data directly (confidentialising);
- deterring any motivation to attempt an identification (e.g. through legally enforceable undertakings not to attempt identification and penalties for breaching the undertakings);
- restricting access;
- educating data users about the importance of protecting privacy, and managing the risks of identification, as well as their obligations in relation to these (e.g. by providing training manuals and detailed instructions); and
- ensuring data is accessed safely through an appropriate environment.

It is not possible to eliminate all identification risks. The objective is to mitigate the risk that an individual or organisation will be identified while maximising the usefulness of the data for statistical and research purposes.

For more information about managing the confidentiality of microdata, or to provide feedback on this series, please email: inquiries@nss.gov.au

Assessing potential identification risks

Assessing microdata for identification risks is a subjective process which requires a detailed examination of the data. Methods to assess identification risk in microdata include:

- cross-tabulation of variables (for example looking at age by income or marital status) to determine unique combinations that may enable a person or an organisation to be identified;
- comparing sample data with population data to determine whether the unique characteristics in the sample are unique in the population; and
- acquiring knowledge of other datasets and publicly available information that could be used for list matching.

The risk of identification can also be assessed by considering factors that contribute to the likelihood of identification (see back page).

Various software packages are available to help assess potential identification risks. These include:

- Mu-ARGUS – a software package developed by Statistics Netherlands. The software is designed to protect against spontaneous recognition only and does not attempt to protect against list matching.
- SUDA – software developed by the University of Manchester. SUDA stands for 'Special Unique Detection Algorithm'. It examines unit record data files and looks for records that are at risk of identification because they have unique combinations of characteristics.

Protecting microdata

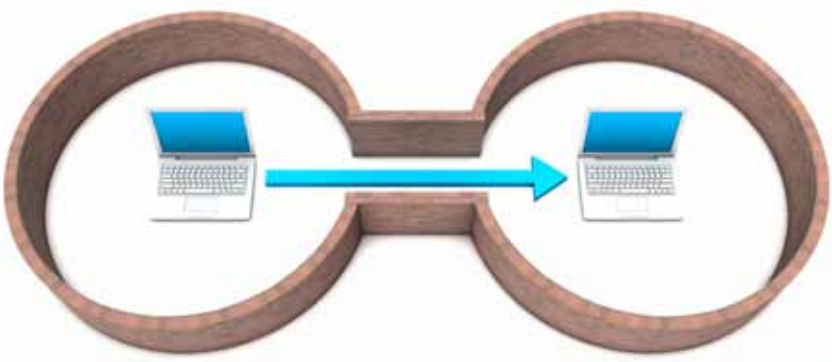
The first level of protection for microdata is to **remove direct identifiers** such as names and Australian Business Numbers. Direct identifiers should always be removed from the data before release. However, there is still a risk of indirect identification occurring in the de-identified data.

Two common approaches to protect microdata are confidentialising and/or restricting access to the file.

Confidentialising microdata

Data perturbation and data reduction methods are used to confidentialise microdata. These are the same basic principles used to protect aggregate data. Popular techniques to confidentialise microdata include:

- limiting the number of variables included in the dataset;
- introducing small amounts of random error (e.g. rounding or data swapping);
- combining categories that are likely to enable identification (e.g. giving age in five year ranges);
- top/bottom coding extreme values of continuous variables like income or age;
- suppressing particular values or records that cannot otherwise be protected from the risk of identification; and
- data swapping – this involves swapping a value in an identifiable record with a value in another record with similar characteristics to hide the uniqueness of the record. For example, a record with a unique language spoken in the region could be swapped with a similar record (based on age, sex, income etc.) in another region where the language is more commonly spoken.



Confidentiality – managing the risk of disclosure in the release of microdata

Restricting access to the microdata file

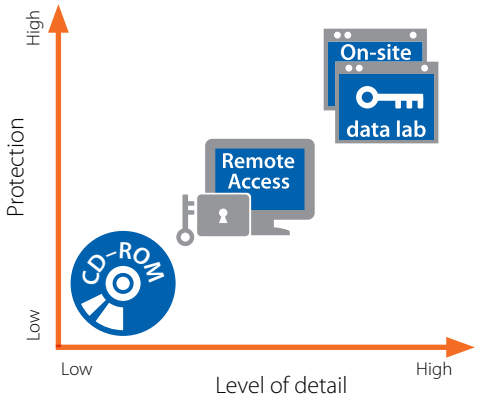
Providing controlled access to microdata is important in protecting the data from identification (or disclosure) risk.

Access to detailed microdata should only be provided under the strictest conditions to approved researchers for an approved purpose. Generally researchers must sign undertakings to abide by specified conditions for access and use of the data.

The extent to which the files are confidentialised will determine how the files are accessed. The more detailed the information, the more protection is required when providing access to microdata.

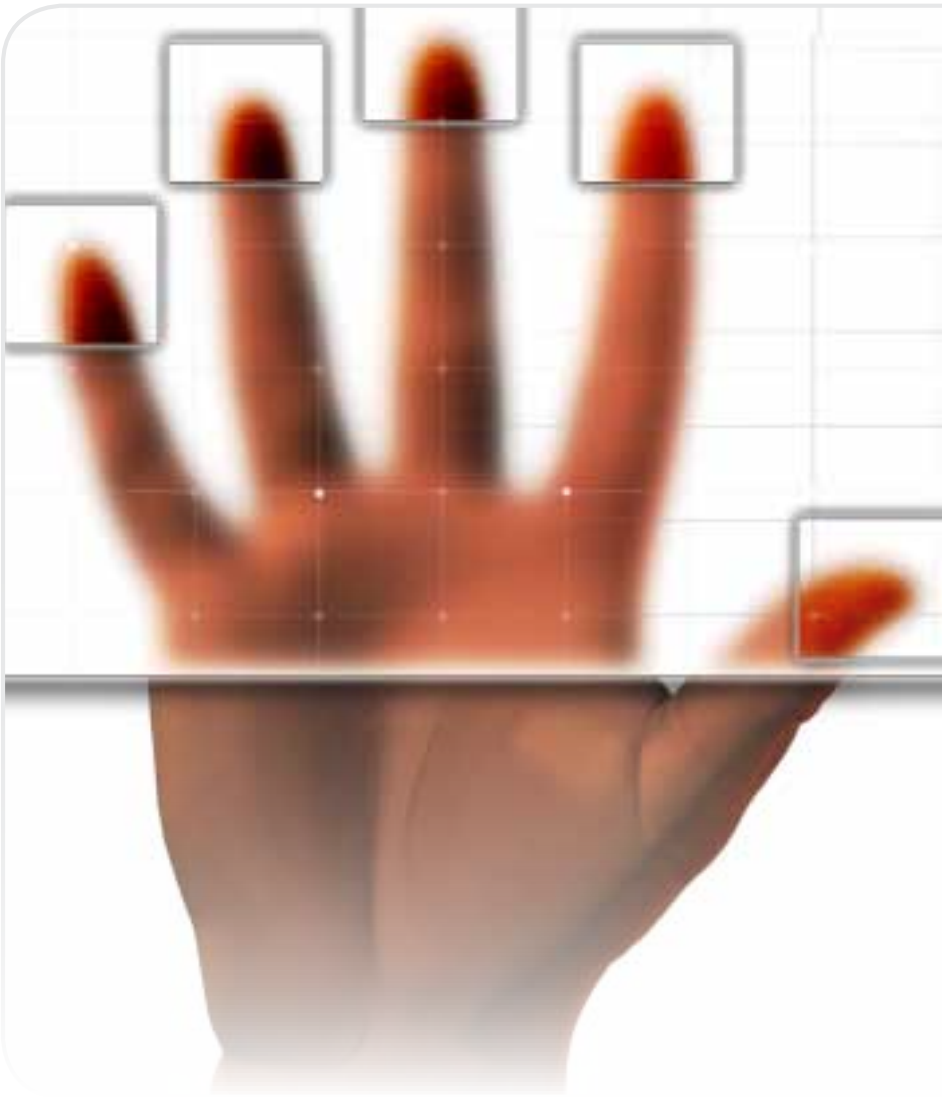
One way of releasing microdata is in the form of Confidentialised Unit Record Files (CURFs). These are files that have been confidentialised to ensure that the direct and indirect identification of individuals or organisations is highly unlikely.

A highly confidentialised microdata file, such as a CURF, may be released publicly on CD-ROM. However, if more detail is left in the CURF, more secure ways of accessing the data need to be used.



Researchers who need lots of detail may have to access the data through a very secure environment. Secure on-site data laboratories are one way of achieving this. The microdata are de-identified and should have some level of confidentialisation to avoid spontaneous recognition, but may still contain data that would allow indirect identification. For this reason, access to this data is available only at data custodian access sites so that all output generated is confidentialised before it leaves the premises.

Another option is providing access to microdata through a **remote access** facility. Remote access facilities are used by statistical agencies and research organisations around the world and mean data users can access microdata from their desktop. Approved researchers can submit data queries through a secure internet-based interface. Requests are generally run against the microdata which is securely stored within the data custodian's computing environment. The results of the queries are confidentialised.



Confidentiality – managing the risk of disclosure in the release of microdata

Factors that increase the risk of identification

Motivation to attempt identification	Motivation is very hard to assess but one issue to consider is whether an organisation or individual would receive any tangible benefit if identification was made.
Level of detail disclosed by the data	The more detail that is included in a unit record, the more likely identification becomes. Data items containing detailed categories, or unit records containing a large number of data items, could reveal enough information to enable identification to be made through a unique combination of characteristics.
Presence of rare characteristics in the data	Even if the number of data items and the number of categories within the data items are limited, there may be a risk of identification if there is a rare and remarkable characteristic in the data. The identification risk posed by the presence of a rare characteristic (or combination of characteristics) depends how remarkable the characteristic is. For example, a 19 year old girl who is widowed is likely to be noticeable in the data and further action would be needed to ensure confidentiality.
Accuracy of the data	The more accurate the dataset, the higher the risk of identification. Conversely, data that are subject to reporting errors, or contain some data items that are not maintained and consequently are out of date, decrease the likelihood of identification.
Age of the data	As a general rule, disclosing an individual's area of residence or marital status 10 years ago is less likely to enable a successful identification than disclosing their current characteristics. However, in some circumstances, the risk of identification can increase as age increases – for example, current quarter earnings for a business may be kept confidential and known only to a small number of people, but previous year earnings may be published in an annual report and so may be publicly available information, enabling identification.
Coverage of the data (completeness)	Complete coverage of a dataset increases the risk of disclosure because a researcher knows that any individual they are aware of in a sub-population will be represented somewhere in the dataset. The sampling process (and the fact that the specific sample selections that have been made are kept confidential) provides some protection against identification.
Presence of other information that can assist in identification	A de-identified dataset cannot, of itself, lead to an identification being made. For an identification to occur a researcher must, implicitly or explicitly, match the data to some other data source, either publicly available information or personal knowledge. Given this, other information that is likely to be accessible or known to the researcher(s) must be assessed.